Correction of Samplable Additive Errors

Kenji Yasunaga
Kanazawa University
Kakuma-machi, Kanazawa, 920-1192, Japan
Email: yasunaga@se.kanazawa-u.ac.jp

***Abstract*—We study the correctability of efficiently samplable errors. Specifically, we consider samplable additive-error channels, where unbounded-weight errors are sampled by a polynomial-time algorithm, and added to the channel input in an oblivious way. Assuming the existence of one-way functions, there are samplable distributions Z over $\{0,1\}^n$ with entropy n^ϵ for $0 < \epsilon < 1$ that are not correctable by efficient coding schemes. Next, we show that there is an oracle relative to which there is a samplable Z with entropy $\omega(\log n)$ that is not correctable by efficient syndrome decoding. For flat distributions Z with entropy m, we show that if Z forms a linear subspace, there is a linear code that corrects Z with rate $R \leq 1 - m/n$. For general flat distributions Z, there is a linear code that corrects Z with error ϵ for rate $R \leq 1 - (m + O(\log(1/\epsilon)))/n$, and no coding scheme can correct Z with error ϵ for rate $R > 1-(m+\log(1-\epsilon))/n$. Finally, we observe that small-biased distributions are not correctable by high-rate codes, and hence there is a small-biased Z with entropy m that is not correctable for rate $R > 1-m/n+(2\log n+O(1))/n$. To derive these results, we use relations between error-correcting codes and other notions such as data compression and randomness condensers.**

I. Introduction

In the theory of error-correcting codes, two of the well-studied channel models are probabilistic channels and worst-case channels. In probabilistic channels, errors are introduced through stochastic processes, and the most well-known one is the binary symmetric channel. In worst-case (or adversarial) channels, we consider the worst-case error for a given error-correcting code and a channel input under the restriction on the weight of the error vector. If we view the introduction of errors as computation of the channel, probabilistic channels perform low-cost computation with little knowledge about the code and the input, while worst-case channels perform high-cost computation with the full-knowledge.

As intermediate channels between these two channels, Lipton [13] considered computationally-bounded channels, where the computation of channels is bounded by polynomial time. Guruswami and Smith [8] presented explicit optimal-rate coding schemes for several computationally-bounded channels, including additive-error channels and time/space bounded channels. For a survey of previous work on intermediate channels, see Section 2 of [9].

In this work, we also focus on computationally-bounded channels. In particular, we consider *additive-error channels*, in which errors are generated independently (or obliviously) of the code and the channel input, and introduced by adding to the input. The binary symmetric channel is an example of additive-error channels. We consider a computationally-bounded analogue of additive-error channels, called *samplable additive-error channels*. In these channels, an error vector is sampled by efficient computation and added to the codeword in an oblivious way. Namely, the sampling algorithm an its random coins cannot depend on the choice of the code or the particular codeword. This is stronger than the standard notion of obliviousness, where an oblivious channel can depend on the code, but not the codeword (cf. [12]). Furthermore, we do not bound the (Hamming) weights of the error vectors. Although most of the existing work considers bounded-weight errors, this restriction may not be necessary for modeling errors generated by nature as a result of polynomial-time computation.

We consider the setting in which coding schemes can be designed with the knowledge on the channel. More precisely, the code can depend on the sampling algorithm of the samplable additive-error channel, but not on its random coins. This setting is incomparable to previous notions of error correction against computationally-bounded channels. Our model is stronger because we do not limit the number of errors introduced by the channel, but is weaker because the error cannot depend on the code or the codeword.

Our Contributions

We would like to characterize samplable additive-error channels regarding the existence of efficient reliable coding schemes. We use the Shannon entropy of samplable distributions as a criterion. The reason is that, if the entropy is zero, it is easy to achieve reliable communication since the error is a fixed string and this information can be used for designing the coding scheme. On the other hand, if the samplable distribution has the full entropy, we could not achieve reliable communication since the truly random error is added to the transmitted codeword. Thus, there seems to be bounds on the existence of efficient reliable coding schemes depending on the entropy of the underlying samplable distribution. When reliable coding schemes exist, an important quantity of the scheme is the *information rate* (or simply *rate)*, which is the ratio of the message length to the codeword length. We investigate the bounds on the rate when reliable communication is achievable.

a) Our Results: Let Z be a distribution over $\{0,1\}^n$ associated with a samplable additive-error channel, and $H(Z)$ the Shannon entropy of Z.

First, we observe that Z can simulate binary symmetric channels. Therefore, it follows from the converse of Shannon's

TABLE I
CORRECTABILITY OF SAMPLABLE ADDITIVE-ERROR Z

$H(Z)$	**Correctabilities**	**Assumptions**	**References**
0	Efficiently correctable	No	Trivial
1	$\forall$ deterministic code, $\exists Z$ not correctable by the code	No	Proposition 2
$O(\log n)$	$\forall$ code of $R > \Omega(\frac{\log n}{n})$ with efficient syndrome decoding, $\exists Z$ not correctable by the code	Oracle access	Theorem 3
$\omega(\log n)$	$\exists Z$ not correctable by efficient syndrome decoding for $R > \omega(\frac{\log n}{n})$	Oracle access	Corollary 1
n^ϵ for $0 < \epsilon < 1$	$\exists Z$ not efficiently correctable	One-way function	Theorem 1
$nH_2(p)$ for $0 < p < 1$	$\exists Z$ not correctable for $R > 1 - H_2(p)$	No	Capacity of BSC
$0 \le m \le n$	$\forall$ linear subspace Z of dimension m is correctable for $R \le 1 - \frac{m}{n}$.	No	Theorem 4
$0 \le m \le n$	$\forall$ flat distribution Z of min-entropy m is		
	(1) correctable with error ϵ for $R \le 1 - \frac{m}{n} - \frac{4\log(1/\epsilon)}{n}$	No	Theorem 5
	(2) not correctable with error ϵ for $R > 1 - \frac{m}{n} + \frac{\log(1/(1-\epsilon))}{n}$	No	Theorem 6
—	$\forall\delta$-biased distribution is not correctable for rate $R > 1 - \frac{2\log(1/\delta)+1}{n}$	No	Theorem 7
$0 \le m \le n$	$\exists$ small-biased Z not correctable for $R > 1 - \frac{m}{n} + \frac{2\log n + O(1)}{n}$	No	Corollary 2
n	Not correctable	No	Trivial

theorem that there exists Z with $H(Z) = n \cdot H_2(p)$ for $0 < p < 1$ that is not correctable by codes with rate $R > 1 - H_2(p)$, where $H_2(p)$ is the binary entropy function defined as $H_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$.

Regarding efficient coding schemes, we observe that if the additive error is pseudorandom (in the cryptographic sense), then no efficient coding scheme can correct the errors. This implies that assuming the existence of one-way functions, there exist Z with $H(Z) = n^\epsilon$ for $0 < \epsilon < 1$ that are not efficiently correctable.

Next, we focus on *linear* coding schemes, where the code forms a linear space. We show that there is an oracle relative to which there exists Z with $H(Z) = \omega(\log n)$ that is not correctable by linear coding schemes that employ efficient *syndrome decoding*. This result also implies that there is no black-box reduction from syndrome decoding algorithms for correcting Z to sampling algorithms for Z. For this result, we use the relation between linear codes correcting additive errors and linear *data compression*.

After that, we consider errors from *flat* distributions Z. When Z forms a linear subspace, we present an efficient coding scheme that corrects Z by syndrome decoding with rate $R \le 1 - m/n$, where $H(Z) = m$. For general flat distributions Z, we show that Z are correctable with error ϵ for rate $R \le 1 - m/n - 4\log(1/\epsilon)/n$. This result is derived by using the relation between a linear code ensemble and a linear *lossless condensers*, established by Cheraghchi [3]. Conversely, we can show that any flat distribution Z is not correctable with error ϵ for rate $R > 1 - m/n + \log(1/(1-\epsilon))/n$. We also observe that no *deterministic* code can correct the family of flat distributions with the same entropy. Specifically, we show that for any deterministic coding scheme, there is a flat distribution Z with $H(Z) = 1$ that is not corrected by the scheme.

Finally, we consider errors from small-biased distributions. We show that δ-biased distributions are not correctable for rate $R > 1 - (2\log(1/\delta) + 1)/n$, and that there is a small-biased distribution Z with $H(Z) = m$ that is not correctable with error ϵ for rate $R > 1 - m/n + (2\log n + O(1))/n$. To derive these results, we use the fact that small-biased distributions can be used for keys of the one-time pad for high-entropy messages, which is derived by Dodis and Smith [6].

The results are summarized in Table I.

b) Related Work: The notion of computationally-bounded channel was introduced by Lipton [13]. He showed that if the sender and the receiver can share secret randomness, then the Shannon capacity can be achieved for this channel. Micali et al. [14] considered a similar channel model in a public-key setting, and gave a coding scheme based on list-decodable codes and digital signature. Guruswami and Smith [8] gave constructions of capacity achieving codes for worst-case additive-error channel and time/space-bounded channels. In their setting of additive-error channel, the weights of errors are bounded, and the errors are only independent of the encoder's random coins. They also gave strong impossibility results for bit-fixing channels, but their results can be applied to channels that use the information on the code and the transmitted codewords. In this work, we give impossibility results even for channels that do no use such information.

Samplable distributions were also studied in the context of data compression [7], [16], [18], randomness extractor [15], [17], [4], and randomness condenser [5].

II. PRELIMINARIES

For a distribution X, we write $x \sim X$ to indicate that x is chosen according to X. We may use X also as a random variable distributed according to X. The *support* of X is $\mathrm{Supp}(X) = \{x : \Pr_X(x) \neq 0\}$, where $\Pr_X(x)$ is the probability that X assigns to x. The *Shannon entropy* of X is $H(X) = E_{x\sim X}[-\log \Pr_X(x)]$. The *min-entropy* of X is given by $\min_{x\in\mathrm{Supp}(X)}\{-\log \Pr_X(x)\}$. For two distributions X and Y defined on the same finite space S, the *statistical distance* between X and Y is given by $\mathsf{SD}(X,Y) = \frac{1}{2}\sum_{s\in S} |\Pr_X(s) - \Pr_Y(s)|$. We say X and Y is ϵ-*close* if $\mathsf{SD}(X,Y) \le \epsilon$. A *flat distribution* is a distribution that is uniform over its support. For $n \in \mathbb{N}$, we write U_n as the uniform distribution over $\{0,1\}^n$.

We define the notion of additive-error correcting codes.

Definition 1: (Additive-error correcting codes) For two functions $\mathsf{Enc} : \mathbb{F}^k \to \mathbb{F}^n$ and $\mathsf{Dec} : \mathbb{F}^n \to \mathbb{F}^k$, and a distribution Z over $\mathbb{F}^n$, where $\mathbb{F}$ is a finite field, we say $(\mathsf{Enc}, \mathsf{Dec})$ *corrects (additive error) Z with error ϵ* if for any $x \in \mathbb{F}^k$, we have that $\Pr_{z\sim Z_n}[\mathsf{Dec}(\mathsf{Enc}(x) + z) \neq x] \leq \epsilon$. The *rate* of $(\mathsf{Enc}, \mathsf{Dec})$ is k/n.

Definition 2: A distribution Z is said to be *correctable with rate R and error ϵ* if there is a pair of functions $(\mathsf{Enc}, \mathsf{Dec})$ of rate R that corrects Z with error ϵ.

We call a pair $(\mathsf{Enc}, \mathsf{Dec})$ a *coding scheme* or simply *code*. The coding scheme is called *efficient* if Enc and Dec can be computed in polynomial-time in n. The code is called *linear* if Enc is a linear mapping, that is, for any $x, y \in \mathbb{F}^n$ and $a, b \in \mathbb{F}$, $\mathsf{Enc}(ax + by) = a\mathsf{Enc}(x) + b\mathsf{Enc}(y)$. If $|\mathbb{F}| = 2$, we may use $\{0, 1\}$ instead of $\mathbb{F}$.

Next, we define syndrome decoding for linear codes.

Definition 3: For a linear code $(\mathsf{Enc}, \mathsf{Dec})$, Dec is said to be a *syndrome decoder* if there is a function Rec such that $\mathsf{Dec}(y) = (y - \mathsf{Rec}(y \cdot H^T)) \cdot G^{-1}$, where $G \in \mathbb{F}^{Rn\times n}$ satisfies that $\mathsf{Enc}(x) = x \cdot G$ for $x \in \mathbb{F}^{Rn}$, and $H \in \mathbb{F}^{n\times Rn}$ is a dual matrix for G (i.e., $GH^T = 0$).

We consider a computationally-bounded analogue of additive-error channels. We introduce the notion of samplable distributions.

Definition 4: A distribution family $Z = \{Z_n\}_{n\in\mathbb{N}}$ is said to be *samplable* if there is a probabilistic polynomial-time algorithm S such that $S(1^n)$ is distributed according to Z_n for every $n \in \mathbb{N}$.

We consider the setting in which coding schemes can depend on the sampling algorithm of Z, but not on its random coins, and Z does not use any information on the coding scheme or transmitted codewords. In this setting, the randomization of coding schemes does not help much.

Proposition 1: Let $(\mathsf{Enc}, \mathsf{Dec})$ be a randomized coding scheme that corrects a distribution Z with error ϵ. Then, there is a deterministic coding scheme that corrects Z with error ϵ.

Proof: Assume that Enc uses at most ℓ-bit randomness. Since $(\mathsf{Enc}, \mathsf{Dec})$ corrects Z with error ϵ, we have that for every $x \in \mathbb{F}^k$, $\Pr_{z\sim Z, r\sim U_\ell}[\mathsf{Dec}(\mathsf{Enc}(x; r) + z) \neq x] \leq \epsilon$. By the averaging argument, for every $x \in \mathbb{F}^k$, there exists $r_x \in \{0, 1\}^\ell$ such that $\Pr_{z\sim Z}[\mathsf{Dec}(\mathsf{Enc}(x; r_x)) \neq x] \leq \epsilon$. Thus, by defining $\mathsf{Enc}'(x) = \mathsf{Enc}(x; r_x)$, the deterministic coding scheme $(\mathsf{Enc}', \mathsf{Dec})$ corrects Z with error ϵ. ■

The fact that the randomization does not help much is contrast to the setting of Guruswami and Smith [8], where the channels can use the information on the coding scheme and transmitted codewords, but not the random coins for encoding. They present a randomized coding scheme with optimal rate $1 - H_2(p)$ for worst-case additive-error channels, for which deterministic coding schemes are only known to achieve rate $1 - H_2(2p)$, where p is the error rate of the channels.

Next, we define the notion of data compression.

Definition 5: For two functions $\mathsf{Com} : \mathbb{F}^* \to \mathbb{F}^*$ and $\mathsf{Decom} : \mathbb{F}^* \to \mathbb{F}^*$, and a distribution Z, we say $(\mathsf{Com}, \mathsf{Decom})$ *compresses Z to length m* if

1) For any $z \in \mathrm{Supp}(Z)$, $\mathsf{Decom}(\mathsf{Com}(z)) = z$, and
2) $E[|\mathsf{Com}(Z)|] \leq m$.

Definition 6: We say a distribution Z is *compressible* to length m, if there are two functions Com and Decom such that $(\mathsf{Com}, \mathsf{Decom})$ compresses Z to length m.

If Com is a linear mapping, $(\mathsf{Com}, \mathsf{Decom})$ is called a *linear* compression.

Finally, we define the notion of lossless condensers.

Definition 7: A function $f : \mathbb{F}^n \times \{0, 1\}^d \to \mathbb{F}^r$ is said to be an *(m, ϵ)-lossless condenser* if for any distribution X of min-entropy m, the distribution $(f(X, Y), Y)$ is ϵ-close to a distribution (Z, U_d) with min-entropy at least $m + d$, where Y is the uniform distribution over $\{0, 1\}^d$. A condenser f is *linear* if for any fixed $z \in \{0, 1\}^d$, any $x, y \in \mathbb{F}^n$ and $a, b \in \mathbb{F}$, $f(ax + by, z) = af(x, z) + bf(y, z)$.

III. Correctability of Additive Errors

A. Errors from Pseudorandom Distributions

We show that no efficient coding scheme can correct pseudorandom distributions.

Theorem 1: Assume that a one-way function exists. Then, for any $0 < \epsilon < 1$, there is a samplable distribution Z over $\{0, 1\}^n$ such that $H(Z) \leq n^\epsilon$ and no polynomial-time algorithms $(\mathsf{Enc}, \mathsf{Dec})$ can correct Z.

Proof: If a one-way function exists, there is a pseudorandom generator $G : \{0, 1\}^{n^\epsilon} \to \{0, 1\}^n$ secure for any polynomial-time algorithm [10]. Then, a distribution $Z = G(U_{n^\epsilon})$ is not correctable by polynomial-time algorithms. If so, we can construct a polynomial-time distinguisher for pseudorandom generator, and the contradiction. ■

B. Uncorrectable Errors from Low-Entropy Distributions

We consider errors from low-entropy distributions that are not correctable by efficient coding schemes. We use the relation between error correction by linear code and data compression by linear compression. The relation was explicitly presented by Caire et al. [2].

Theorem 2 ([2]): For any distribution Z over $\mathbb{F}^n$, Z is correctable with rate R by syndrome decoding if and only if Z is compressible by linear compression to length $n(1 - R)$.

Wee [18] showed that there is an oracle relative to which there is a samplable distribution over $\{0, 1\}^n$ of entropy $O(\log n)$ that cannot be compressed to length less than $n - \Omega(\log n)$ by any efficient compression.

Lemma 1 ([18]): For any k satisfying $6 \log s + O(1) < k < n$, there are a function $f : \{0, 1\}^k \to \{0, 1\}^n$ and an oracle $\mathcal{O}_f$ such that given oracle access to $\mathcal{O}_f$,

1) $f(U_k)$ is samplable, and has the Shannon entropy k.
2) $f(U_k)$ cannot be compressed to length less than $n - 2 \log s - \log n - O(1)$ by oracle circuits of size s.

By combining Lemma 1 and Theorem 2, we obtain the following theorem.

Theorem 3: For any k satisfying $6 \log s + 2 \log n + O(1) < k < n$, there are a function $f : \{0, 1\}^k \to \{0, 1\}^n$ and an oracle $\mathcal{O}_f$ such that given oracle access to $\mathcal{O}_f$,

1) $f(U_k)$ is samplable, and has the Shannon entropy k.
2) $f(U_k)$ is not correctable with rate $R > (2\log s - 3\log n)/n - O(1/n)$ by any linear code $(\mathsf{Enc}, \mathsf{Dec})$ implemented by an oracle circuit of size s, where Dec is a syndrome decoder.

Proof: Item 1 is the same as Lemma 1. We prove Item 2 in the rest. For contradiction, assume that there is an oracle circuit of size s such that the circuit implements a linear code $(\mathsf{Enc}, \mathsf{Dec})$ in which Dec is a syndrome decoder, and $(\mathsf{Enc}, \mathsf{Dec})$ corrects $f(U_k)$ with rate $R > (2\log s - 3\log n)/n + O(1/n)$. By Theorem 2, we can construct $(\mathsf{Com}, \mathsf{Decom})$ that can compress $f(U_k)$ to length $n(1-R) < n - 2\log s - 3\log n - O(1)$, and is implemented by an oracle circuit of size $s + n^2$, where the addition term of n^2 is due to the computation of Com, which is defined as $\mathsf{Com}(z) = z \cdot H^T$. This contradicts Lemma 1. ∎

The following corollary immediately follows.

Corollary 1: For any k satisfying $\omega(\log n) < k < n$, there is an oracle relative to which there is a samplable distribution Z with $H(Z) = k$ that is not correctable by linear codes of rate $R > \omega((\log n)/n)$ with efficient syndrome decoding.

C. Errors from Linear Subspaces

Let $Z = \{z_1, z_2, \ldots, z_\ell\} \subseteq \mathbb{F}^n$ be a set of linearly independent vectors. We can construct a linear code that corrects additive errors from the linear span of Z.

Theorem 4: There is a linear code of rate $1 - \ell/n$ that corrects the linear span of Z by syndrome decoding.

Proof: Consider $n - \ell$ vectors $w_{\ell+1}, \ldots, w_n \in \mathbb{F}^n$ such that the set $\{z_1, z_2, \ldots, z_\ell, w_{\ell+1}, \ldots, w_n\}$ forms a basis of $\mathbb{F}^n$. Then, there is a linear transformation $T : \mathbb{F}^n \to \mathbb{F}^\ell$ such that $T(z_i) = e_i$ and $T(w_i) = 0$, where e_i is the vector with 1 in the i-th position and 0 elsewhere. Let H be the matrix in $\mathbb{F}^{\ell \times n}$ such that $xH^T = T(x)$, and consider a code with parity check matrix H. Let $z = \sum_{i=1}^{\ell} a_i z_i$ be a vector in the linear span of Z, where $a_i \in \mathbb{F}$. Since $z \cdot H^T = (\sum_{i=1}^{\ell} a_i z_i) \cdot H^T = \sum_{i=1}^{\ell} a_i e_i = (a_1, \ldots, a_\ell)$, the code can correct the error z by syndrome decoding. Since $H \in \mathbb{F}^{\ell \times n}$ is the parity check matrix, the rate of the code is $(n - \ell)/n$. ∎

D. Errors from Flat Distributions

Cheraghchi [3] showed a relation between lossless condensers and linear codes correcting additive errors. He gave the equivalence between a *linear* lossless condenser for a *flat* distribution Z and a linear code ensemble in which most of them correct additive errors from Z. Based on his result, we can show that, for any flat distribution, there is a linear code that corrects errors from the distribution.

Theorem 5: For any $\epsilon > 0$ and flat distribution Z over $\{0,1\}^n$ with min-entropy m, there is a linear code of rate $1 - m/n - 4\log(1/\epsilon)/n$ that corrects Z with error ϵ by syndrome decoding.

Proof: Let $f : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^r$ be a linear (m, ϵ)-lossless condenser. Define a code ensemble $\{C_u\}_{u \in \{0,1\}^d}$ such that C_u is a linear code for which a parity check matrix H_u satisfies that for each $x \in \{0,1\}^n$, $x \cdot H_u^T = f(x, u)$. Cheraghchi [3] proved the following lemma.

Lemma 2 (Lemma 15 of [3]): For any flat distribution Z with min-entropy m, at least a $(1 - 2\sqrt{\epsilon})$ fraction of the choices of $u \in \{0,1\}^d$, the code C_u corrects Z with error $\sqrt{\epsilon}$.

We use a linear lossless condenser that can be constructed from a universal hash family consisting of linear functions. This is a generalization of the Leftover Hash Lemma and the proof is given in [3].

Lemma 3 (Lemma 7 of [3]): For every integer $r \leq n, m$, and $\epsilon > 0$ with $r \geq m + 2\log(1/\epsilon)$, there is an explicit (m, ϵ) linear lossless condenser $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^r$.

The statement of the theorem immediately follows by combining Lemmas 2 and 3. ∎

Conversely, we can show that the rate achieved in Theorem 5 is almost optimal.

Theorem 6: Let Z be any flat distribution over $\{0,1\}^n$ with min-entropy m. If a code of rate R corrects Z with error ϵ, then $R \leq 1 - m/n + \log(1/(1-\epsilon))/n$.

Proof: Let $(\mathsf{Enc}, \mathsf{Dec})$ be a code that corrects Z with error ϵ. For $x \in \{0,1\}^{Rn}$, define $D_x = \{y \in \{0,1\}^n : \mathsf{Dec}(y) = x\}$. That is, D_x is the set of inputs that are decoded to x by Dec. Since the code corrects the flat distribution Z with error ϵ, $|D_x| \geq (1-\epsilon)2^m$ for every $x \in \{0,1\}^{Rn}$. Since each D_x is disjoint, $\sum_{x \in \{0,1\}^{Rn}} |D_x| \leq 2^n$. Therefore, we have that $(1-\epsilon)2^m \cdot 2^{Rn} \leq 2^n$, which implies the statement. ∎

By Lemma 2, one may hope to construct a *single* code that corrects errors from any flat distribution with the same entropy, as constructed in [3] for the case of binary symmetric channels by using Justesen's construction [11]. However, it is difficult to achieve since we can show that it is impossible by deterministic coding schemes, and, as presented in Proposition 1, the randomization of coding schemes does not help in our setting.

Proposition 2: For any deterministic code, there is a flat distribution of min-entropy 1 that is not corrected by the code with error $\epsilon < 1/2$.

Proof: Define a flat distribution to be a uniform distribution over two different codewords c_1 and c_2. If the input to the decoder is $c_1 + c_2$, the decoder cannot distinguish the two cases where the transmitted codewords are c_1 and c_2. Thus, the decoder outputs the wrong answer with probability at least $1/2$ for at least one of the two cases. ∎

E. Errors from Small-Biased Distributions

A sample space $S \subseteq \{0,1\}^n$ is said to be δ-*biased* if for any non-zero $\alpha \in \{0,1\}^n$, $|\mathbb{E}_{s \sim U_S}[(-1)^{\alpha \cdot s}]| \leq \delta$, where U_S is the uniform distribution over S. Dodis and Smith [6] proved that small-biased distributions can be used as sources of keys of the one-time pad for high-entropy messages. This result implies that high-rate codes cannot correct errors from small-biased distributions.

Theorem 7: Let S be a δ-biased sample space over $\{0,1\}^n$. If a code of rate R corrects U_S with error $\epsilon < 1/2$, then $R \leq 1 - (2\log(1/\delta) + 1)/n$.

Proof: Assume for contradiction that $(\mathsf{Enc}, \mathsf{Dec})$ corrects $Z = U_S$ with rate R and error ϵ. Dodis and Smith [6] give the one-time pad lemma for high-entropy messages.

Lemma 4: For a δ-biased sample space S, there is a distribution G such that for every distribution M over $\{0,1\}^n$ with min-entropy at least t, $\mathsf{SD}(M \oplus U_S, G) \leq \gamma$ for $\gamma = \delta 2^{(n-t-2)/2}$, where $\oplus$ is the bit-wise exclusive-or.

For $b \in \{0,1\}$, let $M_b \subseteq \{0,1\}^{Rn}$ be the uniform distribution over the set of strings in which the first bit is b. Note that the min-entropy of M_b is $Rn - 1$. Define $D_b = \mathsf{Dec}(\mathsf{Enc}(M_b) \oplus U_S)$. By Lemma 4,

$$\begin{aligned}
&\mathsf{SD}(D_0, D_1) \\
&\leq \mathsf{SD}(\mathsf{Enc}(M_0) \oplus U_S, \mathsf{Enc}(M_1) \oplus U_S) \\
&\leq \mathsf{SD}(\mathsf{Enc}(M_0) \oplus U_S, G) + \mathsf{SD}(G, \mathsf{Enc}(M_1) \oplus U_S) \\
&\leq 2\gamma
\end{aligned}$$

for $\gamma = \delta 2^{(n-Rn-1)/2}$. Since the code corrects U_S with error ϵ, we have that $\mathsf{SD}(D_0, D_1) \geq 1 - 2\epsilon$. Thus, we have that $2\gamma > 1 - 2\epsilon > 0$, and hence $R \leq 1 - (2\log(1/\delta) + 1)/n$. ■

Alon et al. [1] give a construction of δ-biased sample spaces of size $O(n^2/\delta^2)$. This leads to the following corollary.

Corollary 2: There is a small-biased distribution of min-entropy at most m that is not corrected by codes with rate $R > 1 - m/n + (2\log n + O(1))/n$ and error $\epsilon < 1/2$.

IV. Future Work

We present some possible future work of this study.

a) Further study on the correctability: In Theorem 3, we have shown the impossibility result with some restrictions: allowing some oracle access and the decoding is confined to a powerful syndrome decoding. Thus, proving the impossibility results without these restrictions is possible future work.

In this work, we have mostly discussed impossibility results. Thus, showing non-trivial possibility results is interesting, in particular, for the case that the number of errors introduced by the channel is unbounded.

To study the correctability of other samplable additive channels, such as log-space channels and channels computed by constant-depth circuits, may be interesting future work.

b) Generalizing the results of Cheraghchi [3]: In Section III-D, we have used the results of Cheraghchi [3] who showed that a linear lossless condenser for a flat distribution Z is equivalent to a linear code correcting additive error Z. One possible future work is to generalize this result to more general distributions than flat distributions.

Note that the complexity of decoding is not considered in the above equivalence. Namely, lossless condensers do not need to recover the input from the condensed output, and thus the decoder for the corresponding linear codes may not be done efficiently. However, as presented by Cheraghchi [3] for binary-symmetric channels, it is possible to construct an efficient coding scheme using inefficient decoders based on Justesen's concatenated construction [11]. It may be interesting to explore other distributions (or characterize distributions) that can be efficiently correctable by Justesen's construction.

c) Characterizing correctability: We have investigated the correctability of samplable additive errors using the Shannon entropy as a criterion. There may be another better criterion for characterizing the correctability of these errors, which might be related to efficient computability, to which samplability is directly related.

Acknowledgment

This research was supported in part by JSPS Grant-in-Aid for Scientific Research Numbers 23700010, 25106509, and 24240001 We thank anonymous reviewers for their helpful comments.

References

[1] N. Alon, O. Goldreich, J. Håstad, and R. Peralta. Simple construction of almost k-wise independent random variables. *Random Struct. Algorithms*, 3(3):289–304, 1992.

[2] G. Caire, S. Shamai, and S. Verdu. Noiseless data compression with low density parity check codes. In P. Gupta, G. Kramer, and A. J. van Wijngaarden, editors, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 66 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 263–284. American Mathematical Society, 2004.

[3] M. Cheraghchi. Capacity achieving codes from randomness conductors. In *ISIT*, pages 2639–2643. IEEE, 2009. An extended version is available at http://arxiv.org/abs/0901.1866.

[4] A. De and T. Watson. Extractors and lower bounds for locally samplable sources. *TOCT*, 4(1):3, 2012.

[5] Y. Dodis, T. Ristenpart, and S. P. Vadhan. Randomness condensers for efficiently samplable, seed-dependent sources. In R. Cramer, editor, *TCC*, volume 7194 of *Lecture Notes in Computer Science*, pages 618–635. Springer, 2012.

[6] Y. Dodis and A. Smith. Entropic security and the encryption of high entropy messages. In J. Kilian, editor, *TCC*, volume 3378 of *Lecture Notes in Computer Science*, pages 556–577. Springer, 2005.

[7] A. V. Goldberg and M. Sipser. Compression and ranking. *SIAM J. Comput.*, 20(3):524–536, 1991.

[8] V. Guruswami and A. Smith. Codes for computationally simple channels: Explicit constructions with optimal rate. In *FOCS*, pages 723–732. IEEE Computer Society, 2010.

[9] V. Guruswami and A. Smith. Optimal-rate code constructions for computationally simple channels. *CoRR*, abs/1004.4017, 2013. This is an extended version of [8].

[10] J. Håstad, R. Impagliazzo, L. A. Levin, and M. Luby. A pseudorandom generator from any one-way function. *SIAM J. Comput.*, 28(4):1364–1396, 1999.

[11] J. Justesen. Class of constructive asymptotically good algebraic codes. *IEEE Transactions on Information Theory*, 18(5):652–656, 1972.

[12] M. Langberg. Oblivious communication channels and their capacity. *IEEE Transactions on Information Theory*, 54(1):424–429, 2008.

[13] R. J. Lipton. A new approach to information theory. In P. Enjalbert, E. W. Mayr, and K. W. Wagner, editors, *STACS*, volume 775 of *Lecture Notes in Computer Science*, pages 699–708. Springer, 1994.

[14] S. Micali, C. Peikert, M. Sudan, and D. A. Wilson. Optimal error correction for computationally bounded noise. *IEEE Transactions on Information Theory*, 56(11):5673–5680, 2010.

[15] L. Trevisan and S. P. Vadhan. Extracting randomness from samplable distributions. In *FOCS*, pages 32–42. IEEE Computer Society, 2000.

[16] L. Trevisan, S. P. Vadhan, and D. Zuckerman. Compression of samplable sources. *Computational Complexity*, 14(3):186–227, 2005.

[17] E. Viola. Extractors for circuit sources. In R. Ostrovsky, editor, *FOCS*, pages 220–229. IEEE, 2011.

[18] H. Wee. On pseudoentropy versus compressibility. In *IEEE Conference on Computational Complexity*, pages 29–41. IEEE Computer Society, 2004.